프로젝트 보고서

Title : MEPFIL USP 1 값 예측(선형회귀)

과목 : 머신러닝 프로그래밍

지도교수 : 김성수 교수님

작성일 : 2023-06-05

학번 : 2023254002

성명 : 이민수

1. 서 론·····3

1.1 주제 선정·····3

1.2 주제 목적·····3~4

1.1 선형회귀모델의 소개·····4

2. 분석 및 방법·····5

2.1 Flow Chart·····5~7

3. 결과·····8

3.1 상관 관계 분석·····8

3.2 선형 회귀 분석·····9~10

4. 결론·····11

1. 서론

1.1 주제 선정

봉합사는 의료 수술에 필요한 재료로써 고순도 고분자가 필요하다. 또한 안정성이 필요함으로서 염증이나 독성 반응이 없어야 하고 보관이나 소독이 용이해야 한다. 목적에 맞춰 다양한 형태로 가공할 수 있어야 하며 셩체적합성, 생분해성 등 특성을 갖춰야 한다. 봉합사는 생체에 분해 흡수성에 따라 흡수성 봉합사와 비흡수성 봉합사로 나눌 수 있다. 최근에 봉합사는 단순히 조직 접합용 아니라 미용, 성형이나 한방에도 인기를 많이 끌고 있다. 특히 미용 성형 분야에 리프팅실로 많이 사용하고 있다. 리프팅실을 피하조직에 삽입하여 피부를 당기면서 주름 개선 효과를 나타낸다. 그렇기에 봉합사는 인장력과 탄성력 그리고 유연성 모두 좋아야 하며 적절한 분해기간도 필요하다.

(주)메타바이오메드 공정 생산되는 MEPFIL, MEPFIL-LAC, MEPFIL-QUICK, MEPFIL-LAC QUICK, MEPFIL-II 의 총 5 가지 제품이 있다. 생산 조건으로 온도(Temp), 압력(Press), 속도(Speed) 로 설정되며, 그중 인장 강력(gden)은 합격/불합격을 확인하는 수치로써 ㈜메타바이오메드의 회사 수익(불합격 시 불합격 시 전 제품 폐기)과 밀접한 관련이 있는 중요한 요소이다.
이러한 합격/불합격을 향상하거나 향상시키거나 현재의 수율을 유지하기 위해서 ㈜메타바이오메드에서는 다양한 공정 특성을 분석하고 모형화하고 있다. 그러나 생산의 경우 매우 많은 공정변수 및 장비 변수에 장비 변수에 영향을 받아, 합격/불합격 요인을 파악하기 매우 어렵다. 따라서, 단순한 통계적 분석이나 경험적 기술로는 합격/불합격에 따른 수율을 향상하는 데 한계를 지니고 있으며, 다양한 변동 요인을 도출하여 공정 최적화를 이루어 내야 한다.

1.2 주제 목적

본 연구의 목적은 다음과 같다. 첫째, 회귀분석을 활용해 온도(Temp), 압력(Press), 속도(Speed)에서 특성별 중요도를 선정하고 이를 통해 MEPFIL 제품의 어떤 특성이 인장 강력(gden)의 주요한 영향을 미치고 있는지 파악한다. 둘째, 선형회귀분석(linear regression)을 이용하여 MSE(평균제곱오차) 값을 확인하고 Best Model 을 찾는다. 셋째,

결과를 통하여 특성별 중요도가 인장 강력(gden)에 끼치는 영향을 확인한다.
본 연구의 이후 구성은 다음과 같다. 2 장에서는 전체적인 분석 및 방법론에 대해 상세히 서술하였으며 3 장에서는 이를 적용한 결과에 대해 서술하였다. 마지막으로 4 장에서는 결론을 서술하였다.

1.3 선형회귀모델의 소개

통계학에서 선형 회귀(linear regression)는 종속 변수 y와 한 개 이상의 독립 변수 (또는 설명 변수) X와의 선형 상관 관계를 모델링하는 회귀분석 기법이다. 한 개의 설명 변수에 기반한 경우에는 단순 선형 회귀(simple linear regression), 둘 이상의 설명 변수에 기반한 경우에는 다중 선형 회귀라고 한다.
선형 회귀는 선형 예측 함수를 사용해 회귀식을 모델링하며, 알려지지 않은 파라미터는 데이터로부터 추정한다. 이렇게 만들어진 회귀식을 선형 모델이라고 한다.
선형 회귀는 깊이있게 연구되고 널리 사용된 첫 번째 회귀분석 기법이다.[3] 이는 알려지지 않은 파라미터에 대해 선형 관계를 갖는 모델을 세우는 것이, 비선형 관계를 갖는 모델을 세우는 것보다 용이하기 때문이다.
선형 회귀는 여러 사용 사례가 있지만, 대개 아래와 같은 두 가지 분류 중 하나로 요약할 수 있다.

- 값을 예측하는 것이 목적일 경우, 선형 회귀를 사용해 데이터에 적합한 예측 모형을 개발한다. 개발한 선형 회귀식을 사용해 y 가 없는 x 값에 대해 y 를 예측하기 위해 사용할 수 있다.
- 종속 변수 y와 이것과 연관된 독립 변수 X_1, ..., X_p가 존재하는 경우에, 선형 회귀 분석을 사용해 X_j와 y의 관계를 정량화할 수 있다. X_j는 y 와 전혀 관계가 없을 수도 있고, 추가적인 정보를 제공하는 변수일 수도 있다.

2. 분석 및 방법

2.1 Flow Chart

Pig1. Flow Chart 그대로 순서대로 데이터 셋, 변수 선정, 관계도 확인, 모델 선정, Test, Best Model 선정, 결론 순으로 진행을 실시하였다.

Pig1. Flow Chart

2.2.1 제품 선정

제품 선정은 MEPFIL, MEPFIL-LAC, MEPFIL-QUICK, MEPFIL-LAC QUICK, MEPFIL-II 중 regular 한 제품인 MEPFIL 을 선정하였으며, SIZE 는 USP 1 으로 선정하였다. MEPFIL USP 1 은 ㈜메타바이오메드의 주력 상품이며 1 년 중 가장 많이 생산하고, 판매하는 제품이며 인장 강력(gden)의 편차가 그리 많지 않은 점도 있다.

2.2.2 Data 확보

Data 는 2022 년 1 월 1 일 ~ 2022 년 12 월 31 일에 생산된 MEPFIL USP 1 의 제품의 Data 를 확보하였으며, 이를 독립변수, 종속변수로 선정하여. csv File 로 재 정리 진행 하였다.

2.2.3 변수 정의

아래의 Pig 2. MEPFIL USP 1 생산 조건은 아래의 표와 같다. Mult 공정에서 MEPFIL USP 1 을 생산하는데 필요한 조건은 온도(Temp), 압력(Press), 속도(Speed) 임을 확인하였으며, 독립변수로 온도(Temp), 압력(Press), 속도(Speed) 이 세 가지를 선정하였고 종속변수로는 인장 강도(gden)을 선정하였음을 말씀드린다. 마지막으로 위에 언급한 제품 선정, Data 확보, 변수 정의에 따른 파이썬 종류 및 알고리즘을 정리한 내용을 Table1. Data Sampling 으로 표시하였다.

Pig 2. Mult MEPFIL USP 1 조건

그룹 LOT (POP)	TSSR12201G001		방사기	작업일자	시험의뢰	FDY강도	Target	달성율	Denier	1km/g	
			E	2022-01-26 (수)	S2-96	목표	8.4	96%	91.0	양품	229.4
(구) LOT					수율	차이	0.3	4%		core	246.0
제품명(선택)	MEPFIL USP 1				92.3%	투입량	43,492 m		불량수량	3,365 m	

지관 No	섬도 (Denier)	편차 (Denier)	강도 (g/den)	신도 (%)	강력 (gf)	무게 (g)	Chip 이력 사항 LOT / MI / 수분	Chip 이력 사항 사용량(g)	권취 시간	Melt temp (온도℃)	Pack Press (psi)	1 G/R Speed (mpm)	권자 상태 (내,외층)
Avg.	90.7	0.4	8.1	31.3	731			총 9,977 g		261.0	1028	420	
R.	0.8		0.6	4.4	64			↑ TotaL 사용량(g		0.8	30	0	
1	90.5		8.136	30.36	737	1,143	L1-205	9,977	0:40	261.1	1,010	420	양호
2	90.5	0.0	8.130	33.27	735	1,154	22.2/20		1:40	261.1	1,020	420	양호
3	91.0	0.5	8.366	31.52	762	1,151	2021-10-21		2:40	261.1	1,027	420	양호
4	91.0	0.0	7.729	30.49	704	1,151			3:40	261.1	1,025	420	양호
5	91.0	0.0	7.982	31.99	727	1,154			4:40	261.1	1,040	420	양호
6	90.6	0.4	8.317	31.71	753	1,149			5:40	261.2	1,036	420	양호
7	90.2	0.4	7.743	31.87	698	1,150			6:40	261.1	1,033	420	양호
8	90.6	0.4	8.060	28.87	730	1,153			7:45	260.4	1,033	420	양호

Table1. Data Sampling

제품명	MEPFIL USP1
Data 기간	2022 년 1 월 ~ 12 월 생산 Lost
Data 수	3,719EA
파이썬	구글 코랩
선형회귀	Linear, Ridge, Lasso, Elastic Net, RANSAC
X(독립변수)	온도(Temp), 압력(Press), 속도(Speed)
Y(종속변수)	인장강력(Melt gden)

1	Melt gden	Temp	Press	Speed
2				
3	8.136	261.1	1010	420
4	8.130	261.1	1020	420
5	8.366	261.1	1027	420
6	7.729	261.1	1025	420
7	7.982	261.1	1040	420
8	8.317	261.2	1036	420
9	7.743	261.1	1033	420
10	8.060	260.4	1033	420
11	7.726	260.0	1087	495
12	7.766	259.9	1094	495
13	7.879	259.9	1107	495
14	7.582	259.9	1108	495
15	7.635	260.0	1109	495
16	8.287	260.4	1108	490

3701	7.896	269.9	1597	470
3702	8.037	269.8	1603	470
3703	8.373	270.1	1605	470
3704	7.996	270.0	1612	465
3705	8.095	270.0	1610	465
3706	7.901	270.0	1634	465
3707	8.038	270.0	1631	465
3708	7.910	270.0	1590	475
3709	7.688	270.0	1572	475
3710	7.774	269.9	1574	475
3711	7.793	270.0	1575	475
3712	7.726	270.0	1583	475
3713	8.017	270.0	1588	475
3714	8.358	270.0	1600	475
3715	7.688	270.0	1613	475
3716	7.887	270.0	1603	470
3717	7.926	270.0	1604	470
3718	8.085	270.0	1608	470
3719	7.580	270.1	1611	470

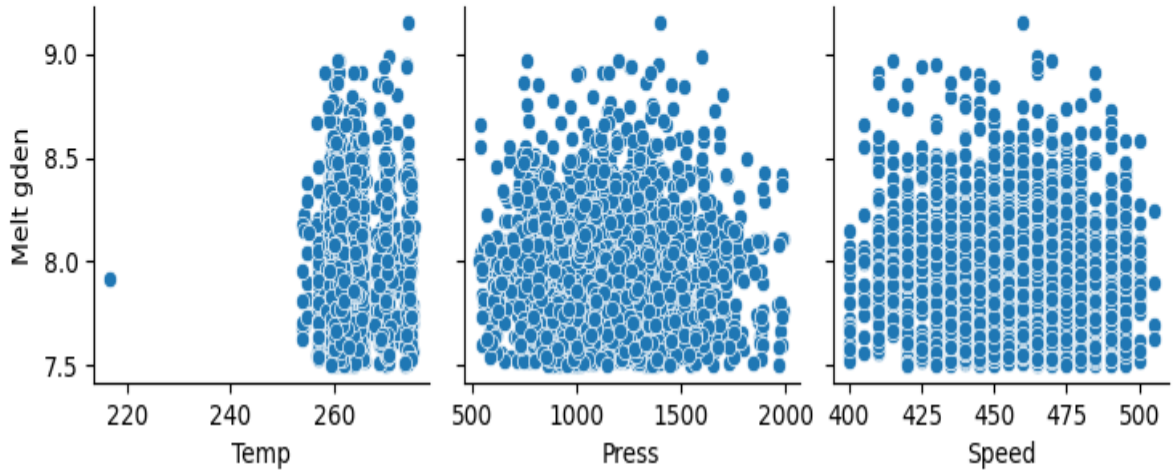
3. 결과

3.1 상관 관계 분석

종속 변수인 인장 강력(gden)을 기준으로 온도(Temp)는 0.002, 압력(Press)은 1.708e-05, 속도(Speed)는 -0.001 의 값을 확인하였다. 결론적으로 온도(Temp), 압력(Press), 속도(Speed) 이 세 가지 독립 변수 중 온도(Temp)가 압력(Prees), 속도(Speed)에 비해 종속 변수인 인장 강력(gden)과 상관관계가 그나마 높다는 것을 확인하였고, 이를 근거로 선형회귀가 아닌 비선형회귀 알고리즘을 사용하기로 결정하였으며, 독립 변수로 더 이상 압력과 속도는 설정하지 않고 온도(Temp)와의 상관관계를 찾는 테스트를 진행하였다.

Fig3. 상관 관계 & 그래프

```
Coefficient for Temp: 0.0022050614808892402
Coefficient for Press: 1.708134945007534e-05
Coefficient for Speed: -0.0014201809196507484
Correlation between Temp and Melt gden: 0.025167487412849406
Correlation between Press and Melt gden: -0.014743259418231621
Correlation between Speed and Melt gden: -0.10519655689041978
```



3.2 선형 회귀 분석

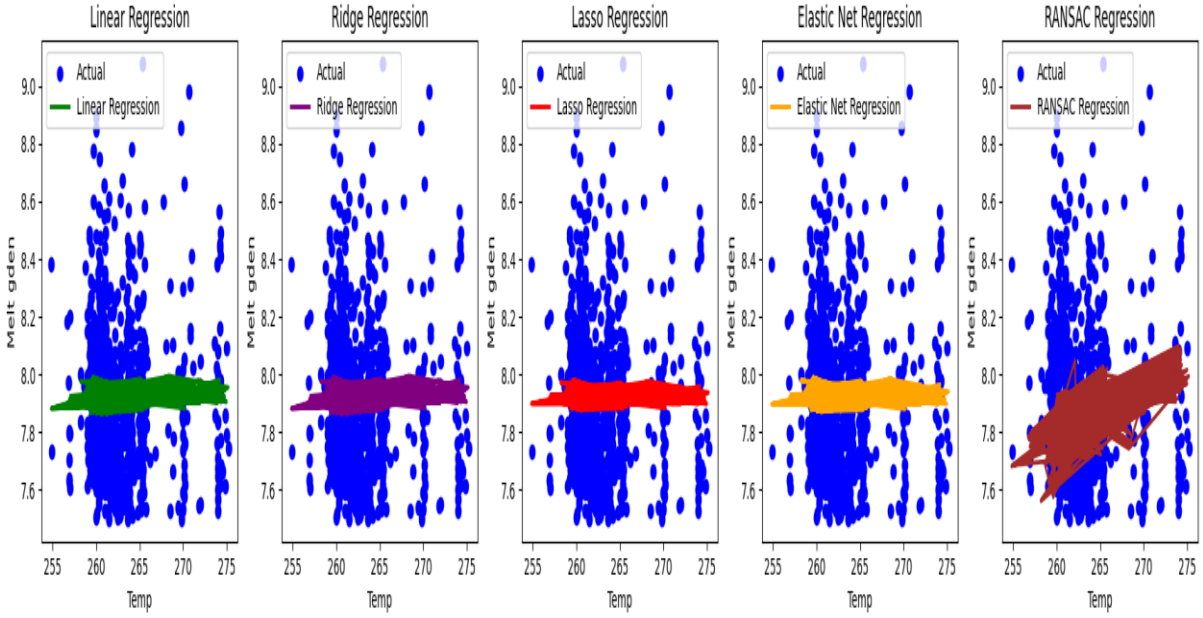
앞서 언급한 Linear, Ridge, Lasso, Elastic Net, RANSAC 이 5 가지 알고리즘을 선정하였으며, 독립 변수로는 온도(Temp)로 선정 후 도출값은 MSE 값 및 Best Model 을 선정하였다. (Pig4. MSE Data & Pig5. Temp-Mult gden 그래프 참조)

Pig4. MSE Data

```
Mean Squared Error:
Linear Regression: 0.0699909595600971
Ridge Regression: 0.0699909598344641 4
Lasso Regression: 0.0701672799827263 7
Elastic Net Regression: 0.0700793866942452 7
RANSAC Regression: 0.0724269097962548 5

Best Model:
Model: Linear Regression
MSE: 0.0699909595600971
```

Pig5. Temp-Mult gden 그래프



Linear, Ridge, Lasso, Elastic Net, RANSAC 모델 중 Best 모델은 Linear 모델임을 확인하였다. 추가로 R2(결정계수) 및 Cross-Validation MSE 값을 확인하였다.
(Fig6. R2&VMSE)

Fig6. R2&VMSE

```
Mean Squared Error (MSE): 0.06999095956000971
R-squared (R2): 0.019752105062800562
Cross-Validated MSE: 0.07107441387897284
```

MSE 값이 낮을수록 모델 성능이 우수함을 나타내며 0.06 의 값은 모델의 성능이 우수함을 나타내고 있다. 다만, R2(결정계수)는 1 에 못미치는 너무 낮은 0.01 로써 독립변수와 종속 변수간 서로 영향을 미치는 관계는 매우 미비하다고 결론 내렸다.

4. 결론

㈜메타바이오메드 MEPFIL USP 1 의 적합/부적합을 나타내는 gden 의 값과 생산 조건으로 온도(Temp), 압력(Press), 속도(Speed)의 관계를 아래와 같이 결론을 내리게 되었다.

- 앞서 R2(상관계수)의 값이 0..01 로써 독립 변수와 종속 변수간 서로의 영향은 매우 미비한 것으로 결론 내렸다. 하지만, 관리 차원에서 작업 표준화에 맞춰 MEPFIL USP 1 생산이 설정된 값을 준수하여 기계를 운용해야 한다.

- 독립 변수 온도(Temp)는 종속 변수 인장 강력(gden) 값에 다른 독립 변수 압력(Press), 속도(Speed)에 비하여 상관 관계가 있지만, 이를 가지고 인장 강력의 값을 추정하기 매우 어려움을 확인 하였다.

- Mult 공정 전 원료 입도 검사(입도분석기로 입도 분석) 및 중합 공정에서의 영향, 그리고 Mult 공정의 Cleanroom 상태 등 다양하고 넓게 모든 공정의 변수 및 특정치를 검토해볼 필요가 있다고 판단 된다.

- 합성고분자인 PGA, PLLA 및 공중합체인 PLGA 을 이용하여 제조하는 생분해성 봉합사는 복잡한 제조 방법 및 그 제조 환경에 따라 인장 강력(gden)의 값을 결정한다.

- 이번 과제를 통하여 비록 상관 관계를 찾지 못하였지만, 졸업 전까지 심도 깊게 다시 한번 자료 조사 및 Data 분석을 통하여 상관 관계를 찾도록 업무 진행 예정이다.